Backpressure in Shared-Memory-Based ATM Switches under Multiplexed Bursty Sources

Fabio M. Chiussi, Ye Xia, and Vijay P. Kumar
AT&T Bell Laboratories, Holmdel, NJ 07733, USA

Abstract

In this paper, we study a shared-memory center-stage switch with input multiplexers and output demultiplexers, where backpressure is applied from the demultiplexers to the center stage, and from the center stage to the multiplexers. We consider three backpressure schemes: i) *Non-Selective Backpressure* (NSB), where a congested center stage or a congested demultiplexer applies backpressure indiscriminately to all the traffic destined to it, regardless of the destination; ii) *Per-Port Selective Backpressure* (PPSB), where the center stage applies backpressure selectively only to the traffic destined to the center-stage port(s) experiencing congestion, but a demultiplexer applies backpressure indiscriminately to all the traffic destined to it; and iii) *Per-Subport Selective Backpressure* (PSSB), where both center-stage switch and demultiplexers apply backpressure selectively only to the traffic destined to the output link(s) experiencing congestion. We show that NSB introduces heavy HOL blocking which limits the throughput of the system and causes heavy losses. On the contrary, PPSB and PSSB achieve high throughputs, and allow to increase buffer utilization in the system while keeping the majority of the buffers physically separate in the input multiplexers, a very desirable feature when implementing systems with large buffer capacity. Both these schemes perform very well in the case where only limited sharing of the buffers in the center stage is allowed, as required to guarantee fairness in the switch. With PSSB, small buffer sizes in the demultiplexers can be used. If the buffers in the demultiplexers are large, PPSB and PSSB offer comparable performance.

1. Introduction

ATM has emerged as the technology of choice for broadband networks. User requirements for current and future generations of switches are evolving rapidly, with constant demand for larger switching capacities, larger buffer capacities, more sophisticated buffer management and congestion control capabilities, higher maximum link rates, and ability to aggregate a growing number, variety, and mix of interfaces to different networks. In addition, as the market arena becomes fiercely competitive, the cost targets for ATM switches are precipitously and steadily dropping.

In this context, switches using shared buffers have become popular [1,2,3], because their reduced buffer requirements [4,5] may translate (at least in a certain range of switching and buffering capacities) into cost advantages. Typically, these switching systems have a hierarchical structure [6], consisting of a center-stage shared-memory switch, with an input and an output stage providing multiplexing/demultiplexing of low-speed links into/from high-speed center stage ports. The hierarchical architecture is necessary in order to connect a single center-stage switch to a large number of relatively low-speeds physical links, and to easily interface a given center-stage switch to different network technologies, so that it can be used in a wide variety of network applications; such a configuration is also efficient, since the cost of a card that accommodates the input or output functionality can be shared among several terminations. In such a shared-memory based-switch, providing large buffer capacities, large switching capacities and sophisticated buffer management may eventually become problematic or costly, due to the centralized approach in buffering and control. The challenge is therefore to distribute buffering and control without compromising the efficiency in buffer utilization obtained by sharing.

As we discuss in this paper, one way to achieve this objective is through the use of backpressure. We study the three-stage hierarchical switch when backpressure is applied from the demultiplexers to the center stage and from the center stage to the multiplexers. We consider both *Non-Selective Backpressure* (NSB), where backpressure is applied indiscriminately to all the traffic destined to a congested center stage or to a congested demultiplexer, regardless of the destination within the center stage or the demultiplexer; and *Selective Backpressure* (SB), where backpressure is applied selectively only to the traffic destined to a congested destination, without affecting the traffic for non-congested destinations. In case of SB, we consider two schemes, namely: i) *Per-Port Selective Backpressure* (PPSB), where a congested center stage applies backpressure selectively only to the traffic destined to the port(s) experiencing congestion, but a congested demultiplexer applies backpressure indiscriminately to all the traffic destined to it; and ii) *Per-Subport Selective Backpressure* (PSSB), where both demultiplexers and center-stage switch apply backpressure selectively only to the traffic destined to the output link(s) experiencing congestion. Using simulation, we study throughput and cell loss rate in the three-stage switch under traffic generated by a large number of bursty virtual connections multiplexed on each link terminated at the switch.

We show that NSB introduces heavy HOL blocking which limits the throughput of the system and induces heavy losses. On the contrary, PPSB and PSSB achieve high throughputs, and allow to increase buffer utilization in the system while keeping the majority of the buffers physically separate in the input multiplexers, a very desirable feature when implementing systems with large buffer capacity. The buffer requirements using these two backpressure schemes are much smaller than those of a hierarchical switch with partitioned buffers in the center stage without backpressure, and not much larger than those of a shared-memory center stage without backpressure. Specifically, PPSB is a way to increase the buffer capacity of the center stage using the buffers in the multiplexers (or, in other words, to achieve "sharing" of the buffers in the center stage and in the multiplexers), with only a small penalty in buffer efficiency with respect to "true" memory sharing. PSSB is a way to use the buffers in the multiplexers and in the center

0743-166X/96 $5.00 © 1996 IEEE

stage as an extension of the buffers in the demultiplexers, without introducing HOL blocking. PSSB is the only scheme that achieves high throughput and low loss rates with small buffers in the demultiplexers. If the buffers in the demultiplexers are large, PPSB and PSSB offer comparable performance.

PPSB and PSSB achieve maximum throughput by using partitioned buffers in the center stage; for given buffer sizes in the three-stage switch, minimum cell loss is achieved by allowing partial sharing of the center-stage buffers; thus, both these schemes perform very well in the case where a constraint is placed on the amount of buffer sharing that is allowed in the center stage, as it is required in practice to guarantee fairness in the switch.

NSB has been suggested in the past in switches with input and output buffers as a way to use the capacity of the input buffers in support of the output buffers; most of these studies have considered only uniform traffic [7,8,9,10,11], and not addressed bursty traffic, a key motivating factor for the use of backpressure. NSB has been studied under bursty traffic in buffered multistage networks in [12,13,14]. The interaction between the demultiplexers and the center-stage switch using NSB has been studied in the context of hierarchical switches in [15]. PPSB was originally proposed in the context of switches with input and output buffers in [16], as a way to relieve the HOL problems introduced by NSB, but never generalized to the three-stage hierarchical configuration, and to the most interesting case of limited sharing of the buffers.

This paper is organized as follows. In Section 2, we describe the switch model, the backpressure schemes, and the traffic model used. In Section 3, we study the interaction of the first two stages when backpressure is applied from the center-stage switch to the input multiplexers, and characterize throughput and buffer requirements for each of the backpressure schemes. In Section 4, we expand the characterization of Section 3, by considering the whole three-stage switch with backpressure from the demultiplexers to the center stage, and from the center stage to the multiplexers.

2. Switch and Traffic Models

2.1. Switch Architecture and Backpressure Schemes

In this paper, we consider the switch model shown in Fig. 1. It consists of an NxN center-stage shared-memory switch, with N input multiplexers, each multiplexing n input links into a single center-stage port, and N output demultiplexers, each demultiplexing a center-stage port into n output links. We refer to as *subports* the input ports of the multiplexers and the output ports of the demultiplexers. Each port of the center-stage NxN shared-memory switch has capacity R; each input subport in the multiplexers and each output subport in the demultiplexers has capacity R/n.

Buffering is provided in each stage. The buffer size per port[1] in the center-stage switch is B_{sw}. As it is well known [1,2,3,4,5], internally to the shared-memory switch in the center stage, cells are organized in separate queues, one queue per destination, and the memory is shared among all queues. The memory in the switch may be fully shared, meaning that a single queue can potentially occupy the whole buffer space, or, as it is typical in practice, a limitation may be placed on the maximum length of each queue, in order to avoid that a subset of queues unfairly hogs the whole memory, thus preventing cells destined to other queues from accessing the buffer, and reducing throughput. Each demultiplexer in the output stage has a buffer of capacity B_{dx} to demultiplex traffic from the high-speed center-stage port into the n low-speed subports. Each demultiplexer organizes the cells into separate queues, one queue per subport, and the buffer is shared among all queues; similarly to the switch memory, the memory may be fully shared, or a limitation may be placed on the maximum space allowed for each queue. (Clearly, in case $n = 1$ the output demultiplexers are not necessary.) The buffer size in each multiplexer in the input stage is B_{mx}. As explained in the following, depending on the backpressure scheme used, in the multiplexers cells may be organized in one or more queues; in the latter case, the memory may be fully shared among the queues or a limitation on the maximum space per queue may be used. (Note that, in the case where no backpressure is used in this architecture, the input multiplexers would require no buffering, since no statistical multiplexing occurs in the input stage.)

We characterize the architecture of Fig. 1 with backpressure applied from the demultiplexers to the shared-memory switch and from the switch to the multiplexers. A demultiplexer applies backpressure to the switch when it has no buffer capacity to accommodate a cell destined to one of its subports. When backpressure is applied by the demultiplexer, the switch stores the cell in its buffer rather than transmitting it to the demultiplexer. Similarly, the switch applies backpressure to an input multiplexer when it has no buffer capacity to accept a cell destined to one of its ports, and the cell is

[1] Throughout this paper, buffer sizes are given in number of cells.

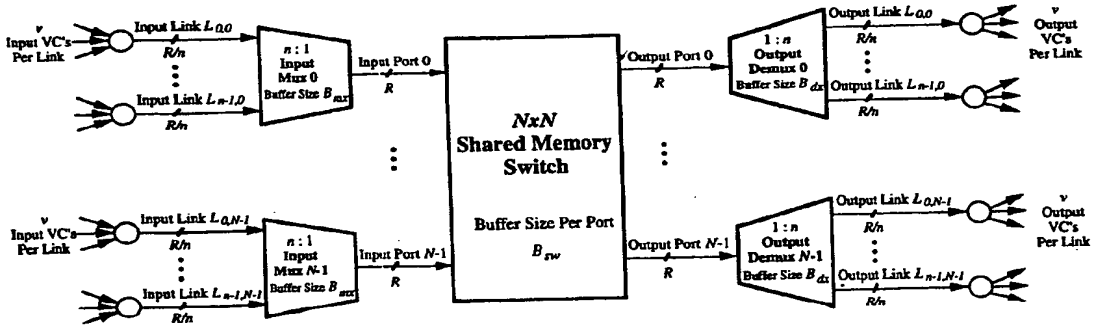

Fig. 1. Switch model.

stored in the multiplexer's buffer rather than being delivered to the switch. With backpressure, cell loss occurs only in the multiplexers (contrary to the case without backpressure, where cell loss occurs in the switch and in the demultiplexers). We study three backpressure schemes.

In *Non-Selective Backpressure* (NSB), a demultiplexer that has no buffer capacity to accept a cell destined to a subport (*i.e.*, either because the demultiplexer's buffer is full, or because the queue for that subport has used all its allowed buffer space), applies backpressure to *all* the cells in the center-stage switch destined to that demultiplexer, regardless to which of the subports the cells are destined. Similarly, the center-stage switch, once it does not have buffer capacity for a cell destined to one of its ports (*i.e.*, either because the switch buffer is full, or because the queue for that port has used all its allowed buffer space), applies backpressure to *all* the cells in the multiplexers destined to the switch, regardless to which ports the cells are destined.

In *Selective Backpressure* (SB), backpressure is applied only to the traffic flow destined to the specific destination(s) experiencing congestion. We consider two types of selective backpressure. In *Per-Port Selective Backpressure* (PPSB), a demultiplexer that has no available buffer space applies backpressure to the center-stage switch in the same way as in non-selective backpressure (*i.e.*, indiscriminately, to all cells in the switch destined to the demultiplexer). The center-stage switch, however, once it cannot accept a cell destined to one of its output ports because of lack of buffer space, applies backpressure *only* to the cells in the multiplexers destined to that specific port, without stopping cells destined to other ports for which there is still available buffer capacity in the switch. In order to implement PPSB, in the multiplexers cells must be sorted in separate queues according to their switch port destination (*i.e.*, each multiplexers maintains N queues). We assume that each multiplexer serves its queues in a round-robin fashion, and visits the queues until it finds one that has a cell that can be delivered to the switch; then, in the following cell time, it resumes visiting the queues from the next queue in the round-robin sequence. When the service mechanism in the multiplexers visits a queue, it checks if the queue is active (*i.e.*, if it has one or more cells) and if backpressure is not applied for that queue; if both conditions are met, the multiplexer delivers the first cell in the queue to the switch, otherwise it visits the next queue. The service mechanism in the multiplexers is sufficiently fast to be able to visit all the queues within one cell time, so that a cell is always delivered from each multiplexer to the switch every cell time, unless all queues are empty or blocked because of backpressure.

In *Per-Subport Selective Backpressure* (PSSB), a demultiplexer that has no buffer space to accept a cell for a certain subport applies backpressure *only* to the cells in the center-stage switch destined to that specific subport, without stopping cells destined to other subports for which buffer space is still available. In order to implement PSSB, cells in the center-stage switch must be organized in separate queues, one per each subport in each demultiplexers (*i.e.*, the switch maintains Nn queues). In the switch, we assume that the n queues corresponding to each port are visited in a round robin fashion, and that the service mechanism is able to visit all the queues within one cell time, so that a cell is always delivered from the switch to each demultiplexer every cell time, unless all queues are empty or blocked. The center-stage switch, in turn, once it has no buffer space to accept a cell for a certain subport in one of the demultiplexers, applies backpressure *only* to cells in the multiplexers destined to that subport. In the multiplexers, as it is the case in the center-stage switch, cells must be organized in separate queues, one per each subport in the system (*i.e.*, each multiplexer also maintains Nn queues). Similarly to PPSB, we assume that the queues in each multiplexer are served in a round robin-fashion[2], and the service mechanism is able to visit all the queues within one cell time (*i.e.*, the only difference with respect to PPSB is the larger number of queues that have to be visited[3]).

In PPSB, each multiplexer actually performs *switching*, since it sorts the cells according to the output *port* to which they are destined. Since in each multiplexer N queues have to be maintained and visited in each cell time to support PPSB, the hardware complexity of the multiplexer is obviously higher than in the case of NSB, where only a single FIFO per multiplexer is sufficient. However, although the buffer management is more complex, the memory bandwidth in the multiplexer is the same as in NSB. To support PPSB, the center stage must send backpressure information to each multiplexer identifying whether or not each of its N queues may receive cells (for example, this may be accomplished by sending N bits from the center stage to each multiplexers), rather than simply informing the multiplexer of whether there is buffer space available, as it is sufficient in NSB (which can be accomplished by transmitting a single bit). In PSSB, each multiplexer performs *switching* to the output *subports*, and has to maintain and visit Nn queues in each cell time. The memory bandwidth is the same as in the other schemes. The backpressure information from the center stage to each multiplexer must consist of the status of Nn queues. The center stage is also more complex than in the other schemes, since it performs switching to the subports. The center stage must receive information of buffer availability of each of the n queues in each of the N demultiplexers, rather than simply availability status of the buffers in the demultiplexers as in the other schemes. Clearly, the relative complexity of the three schemes depends on the specific values of n and N used. For example, with typical values such as $N = 16$ and $n = 16$, PPSB does not represent an implementation challenge with current technology. On the contrary, the implementation of PPSB may be significantly more complex, especially due to the required speed of the mechanism that visits the large number of queues, and to the amount of backpressure information that has to be transmitted between the stages.

From this discussion, it is obvious to expect that the performance of the three backpressure schemes depends heavily on whether the memory in the demultiplexers and center-stage switch is fully shared, or, as it is common in practice, a limitation is placed on the maximum length of each queue in order to guarantee fairness, since this defines

[2] More precisely, in case of PSSB, we assume that in each multiplexer the n queues corresponding to each output port are grouped together; round robin is performed hierarchically, first among the groups of queues, and then within each group (so that in consecutive cell times, if there are cells, queues corresponding to different output ports are served).

[3] Throughout this paper, we consider the case where, in the center-stage switch and in each multiplexer and demultiplexer, all cells destined to a given destination (*i.e.*, a port or a subport, depending on the backpressure scheme) are kept in a *single* queue. This model can be easily generalized to a case where cells for a given destination are further separated in multiple queues (as it is the case, for example, when traffic with different delay priorities must be supported in the switch).

how much buffer space is available for a certain destination at a given time, and ultimately determines when backpressure is triggered. Different limitation mechanisms for the queue length have been conceived. One possibility is to introduce a *threshold* which defines the maximum length of a queue, above which cells can not be accepted for that queue; the threshold could be statically assigned or dynamically changed according to the buffer occupation [17,18]. Alternatively, a *minimum space* can be reserved for each queue, while the rest of the buffer is fully shared. In this paper, we are interested in characterizing the general behavior of the system with the backpressure schemes described above, rather than exploring the details of the different limitation mechanisms on the queue length, which only impact the degree of sharing that can be achieved in the buffers. For this reason, the simplest (and most popular) method of statically-defined thresholds to limit the length of an individual queue in the shared buffers is sufficient for our purposes, and is used throughout our discussion. Specifically, we introduce a static threshold T_{sw} in the center-stage switch, and study the system for different values of T_{sw}, from the case of completely-partitioned buffers (*i.e.*, $T_{sw} = B_{sw}$ for PPSB, or $T_{sw} = B_{sw}/n$ for PSSB), up to the case of no limitation on the queue length (*i.e.*, $T_{sw} \geq NB_{sw}$). The general conclusions that we draw using static thresholds can be easily generalized when other queue-length limitation mechanisms are used.

2.2. *Traffic Model*

We characterize the performance of the switch model depicted in Fig. 1 under bursty traffic generated by a number v of virtual connections multiplexed on each input link. (Here, all virtual connections are unicast connections.) Let ρ be the desired traffic load on the center-stage NxN switch. The traffic on each virtual connection is modeled as the output of an on-off source, which generates bursts of geometrically-distributed lengths, with average burst length L. Individual virtual connections have a peak rate W_{peak} equal to the link rate (*i.e.*, $W_{peak} = R/n$), and a load $\rho_{vc} = W_{ave} / W_{peak} = \rho/v$, where W_{ave} is the average bandwidth of the connection. (The virtual connections are multiplexed on each link in a way that generates the most stressful traffic conditions offered to the switch, namely each time a source generates a burst, it transmits the full burst on the link at the link rate.) All virtual connections are homogeneous.

The traffic offered to the switch is generated as follows. First, given a desired switch load ρ, ρ_{vc} is chosen so that v is a sufficiently large number (*e.g.*, for most of the following discussion, we study ρ equal to 80% and 99%; ρ_{vc} is chosen equal to 1%). A traffic matrix that specifies source and destination of each virtual connection is then randomly generated. For each virtual connection multiplexed on each input link, its destination is randomly chosen, uniformly over all possible destinations. Destinations are assigned sequentially and in a round robin fashion over all input links of the switch until v virtual connections are multiplexed on each output link of the switch. (In other words, the output link bandwidth is allocated according to the average bandwidth of the connections, until all links are uniformly loaded.) Once the traffic matrix is generated and applied to the switch, the on-off sources are started, and the system is simulated for a simulation time sufficiently long to stabilize the queues. The procedure is then repeated, and the system is simulated over a number of random traffic matrices sufficiently large to make the results statistically significant[4]. This traffic model approximates a realistic scenario where a large number of bursty virtual connections are handled by the switch, and the duration of the connections is long.

3. Backpressure in Two-Stage Switch Architecture

In this section, we concentrate on the interaction between center-stage switch and multiplexers. For this purpose, we study a special case of the switch model described above where the demultiplexers are not present (as shown for convenience in Fig. 2). This configuration models the case where $n = 1$; it is also useful to study the switch architecture with $n > 1$, when backpressure is used only between the center-stage switch and the multiplexers. In this latter system, the first two stages can be studied separately, and the demultiplexers behave in the same way as in the case without backpressure, at least for low cell loss rates (in other words, this is the case when backpressure is only used to increase the switching capacity of the center stage, while the demultiplexers are independently sized as in the case without backpressure). In Section 4, we consider backpressure from the demultiplexers, and study the whole three-stage configuration.

Throughout this paper, we assume $N = 8$, and consider $n = 1$, 4, 12, and 25 (*i.e.*, with port rate $R = 622.08$ Mbps, these n's would correspond to link rates of 622.08 Mbps, 155.52 Mbps, 51.84 Mbps, and about 25 Mbps). The load ρ_{vc} of each virtual connection is assumed to be 1%, to make v sufficiently large. The average burst length L in our simulation studies is 10 cells. It is well known, however, that in a shared-memory switch the buffer requirements scale approximately linearly with the burst size [16], and we have confirmed that this rule of thumb holds rather well also in the switch model used in this paper, for all the backpressure schemes. For this reason, we present our results with all the buffer sizes normalized to the burst size, since they actually hold independently from the burst size.

[4] Note that, in the case without backpressure, because of the way the traffic matrix is generated, the traffic in the system would depend only marginally on the traffic matrix. In the case with backpressure, the traffic in the system depends on the interactions between the various stages, and thus it depends more markedly on the traffic matrix.

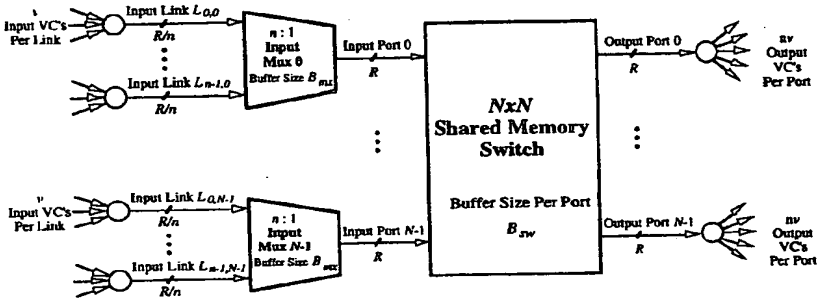

Fig. 2. Two-stage switch model.

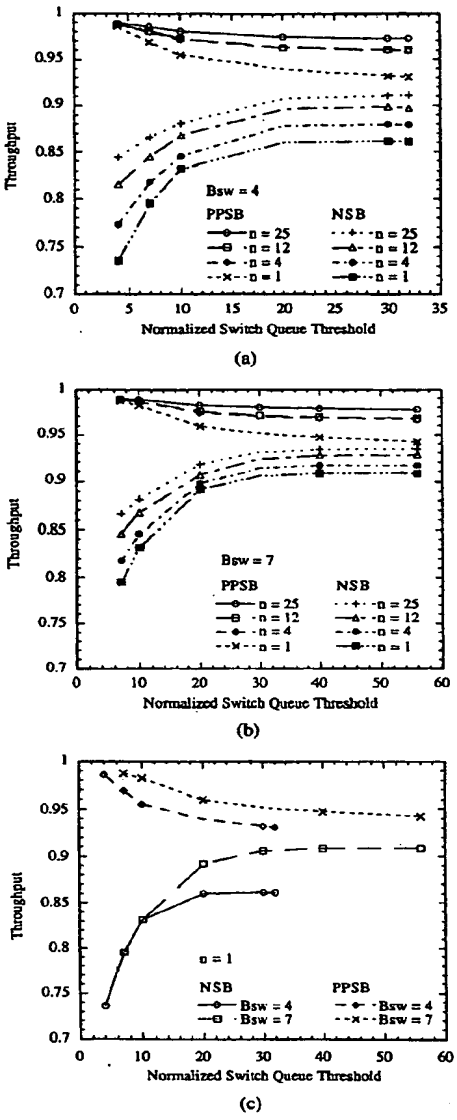

Fig. 3. Throughput of the two-stage switch architecture using *Non-Selective Backpressure* (NSB) and *Per-Port Selective Backpressure* (PPSB), as a function of the normalized queue threshold in the switch T_{sw}, N = 8, ρ = 99%, ρ_{sc} = 1%; the multiplexers have infinite buffers. a) Normalized buffer size per port in the switch $B_{sw} = 4$, for various n; b) $B_{sw} = 7$, for various n; c) $n = 1$, $B_{sw} = 4$ and $B_{sw} = 7$.

3.1. Throughput

In this subsection, we show that the maximum throughput of the switch model in Fig. 2 heavily depends on the backpressure scheme used. Since there are no demultiplexers, the only backpressure schemes that are relevant in this case are NSB and PPSB. In our study, we simulate the switch model under a traffic load on the center-stage switch approaching 100% (the results presented here are for ρ = 99%), with infinite buffers in the multiplexers, for different sizes

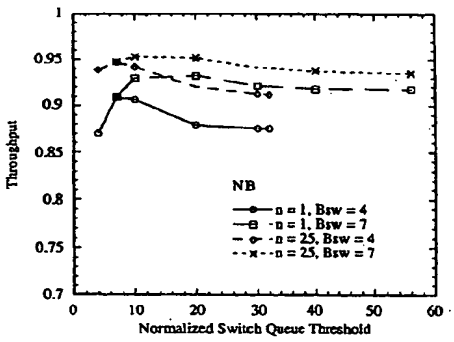

Fig. 4. Throughput of the two-stage switch architecture with *No Backpressure* (NB), for $B_{sw} = 4$ and $B_{sw} = 7$, as a function of T_{sw}, for $n = 1$ and $n = 25$; $N = 8$; $\rho = 99\%$, $\rho_{sc} = 1\%$.

of the switch buffer. Indeed, the throughput of such a system inherently depends on the buffer size in the center-stage switch, which is what determines how often backpressure is applied; thus, the throughput must be characterized with finite buffers in the switch.

The throughput of the two-stage switch using NSB and PPSB as a function of the normalized queue threshold[5] in the switch T_{sw} is shown in Fig. 3(a) and 3(b), for normalized buffer size per port in the switch $B_{sw} = 4$ and $B_{sw} = 7$, for various n. The throughput with NSB is always well below 1. This is due to Head-Of-Line (HOL) blocking induced by non-selective backpressure, since a cell that is blocked at the head of the queue in a multiplexer due to backpressure, applied because of lack of buffer space in the center stage, may block cells for other ports which could be still accepted in the buffers in the center stage, but have instead to wait unnecessarily. The effect is especially marked for the case of partitioned buffers in the center-stage switch (*i.e.*, $T_{sw} = B_{sw}$), since this is the case when backpressure is applied more often; the throughput degrades as low as 72% for n = 1 and $B_{sw} = 4$.

As the threshold increases, sharing of the buffer increases, improving buffer utilization, and backpressure is applied less often; HOL blocking is therefore reduced and the throughput improves monotonically until the buffer is fully shared (*i.e.*, $T_{sw} = NB_{sw}$): In any case, however, the throughput remains below 1, due to the finite buffer size of the center-stage buffer. For any given value of T_{sw}, the throughput improves as n increases. With larger n, the multiplexing effect on the traffic arriving at the multiplexers on several links smoothes the burstiness of the traffic offered to the center-stage switch, thus reducing the buffer requirements in the switch; hence, for a certain size of the switch buffer, backpressure is applied less often. The throughput also increases as the buffer size in the switch increases (see Fig. 3(c)), again because the larger buffer size translates to backpressure applied less often.

It should be noted that the throughput is always below the throughput of a shared-memory switch of the same size with *No Backpressure* (NB, shown in Fig. 4), even in the best case of full sharing of the switch buffer; this confirms the well known fact (seen for example in a crossbar with input buffers) that discarding cells

[5] The threshold values have been normalized to the burst length, to make them consistent with the normalized buffer sizes.

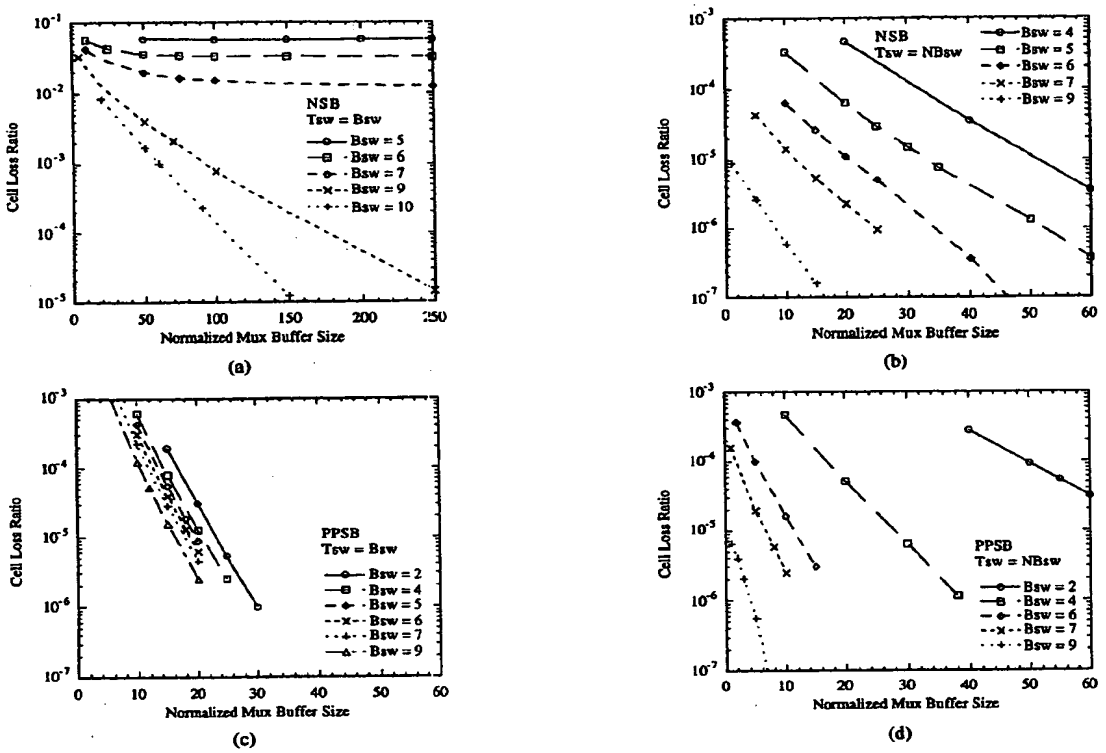

Fig. 5. Cell loss rate vs. normalized buffer size in each multiplexer B_{mx} using NSB and PPSB, for different B_{sw}, $n = 1$, $N = 8$; $\rho = 80\%$, $\rho_{uc} = 1\%$. a) NSB, $T_{sw} = B_{sw}$; b) NSB, $T_{sw} = NB_{sw}$; c) PPSB, $T_{sw} = B_{sw}$; d) PPSB, $T_{sw} = NB_{sw}$.

that cannot proceed to their destination gives higher throughput than storing cells in input buffers, if storing induces HOL blocking. Note that, with no backpressure, the throughput of the system would tend to 1 as the buffer size in the switch increases; with NSB, the throughput is limited, despite the infinite buffers in the multiplexers.

The behavior is quite different with PPSB. By sorting the traffic per output port, and applying backpressure selectively only to the traffic destined to a congested port, this scheme does *not* introduce HOL blocking. Indeed, the maximum throughput is 99%, equal to the offered load, for both $B_{sw} = 4$ and $B_{sw} = 7$. It is interesting to note that the maximum throughput is achieved with completely partitioned buffers in the center stage, and monotonically decreases as the threshold increases, thus with a trend opposite to the one of NSB (also, the maximum throughput in PPSB does not depend on the buffer size in the switch, contrary to NSB); the degradation in throughput is slower as the switch size increase, as shown in Fig. 3(c). The lowest throughput occurs in the case of fully-shared buffers, weakly depends on the buffers size, and is always above 90%. The monotonic decrease in throughput in PPSB is due to the fact that, as T_{sw} becomes greater than B_{sw}, there is a chance for a subset of queues to hog the whole available buffer space in the center-stage switch, thus affecting the throughput, and the effect is more and more likely as the switch threshold increases. This effect can be seen in the system without backpressure (see Fig. 4), where, as the threshold increases from B_{sw}, the throughput initially increases rapidly, because sharing of the memory improves; the throughput, however, soon reaches a peak for a value of the threshold well below NB_{sw} (the throughput at the peak obviously depends on the switch buffer size) and then slowly decreases, as the likelihood of memory hogging increases.

For the same B_{sw}, the throughput of PPSB is always well above the throughput of NSB. It is surprising to note that this holds even for fully-shared buffers in the center stage, a case for which PPSB appears very similar to NSB; PPSB retains its advantage over NSB due to the smoothing effect on the traffic offered to the center-stage switch introduced by the round-robin service of the different queues in the multiplexers [19] (in NSB, there is a single queue in each multiplexers and cells are served in FIFO order).

3.2. Cell Loss Rate

In this section, we study the buffer requirements in the center-stage switch and in the multiplexers, for different values of the switch threshold T_{sw}. The results presented here are for a load $\rho = 80\%$, $\rho_{uc} = 1\%$; We assume fully-shared buffers in the multiplexers. We consider the case $n = 1$; similar results can be obtained for other n's.

In Fig. 5, we show the cell loss rate using NSB and PPSB vs. the normalized buffer size in each multiplexer B_{mx}, for different B_{sw}, for $T_{sw} = B_{sw}$ and $T_{sw} = NB_{sw}$; in Fig. 6, we show the normalized buffer requirements per port in the switch and in the multiplexers to achieve 10^{-6} cell loss rate using the two backpressure schemes, for different T_{sw}. To serve as a reference, in Fig. 7 we plot the cell loss rate in the system with *No Backpressure* vs. the normalized buffer size per port in the switch.

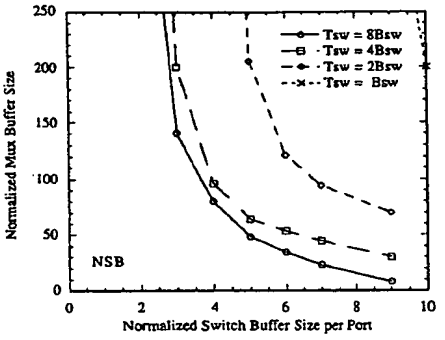

Fig. 6. Normalized buffer requirements per port in the switch and in each multiplexer to achieve 10^{-6} cell loss rate, for different normalized queue thresholds in the switch T_{sw}, $n = 1$, $N = 8$; $\rho = 80\%$, $\rho_{uc} = 1\%$. a) NSB; b) PPSB.

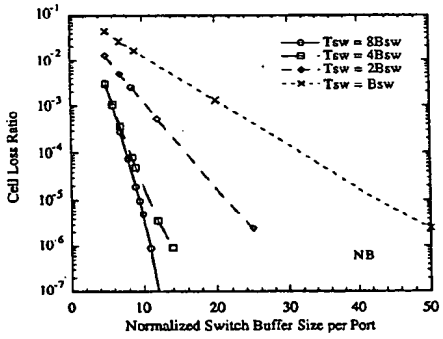

Fig. 7. Cell loss rate in the switch with *No Backpressure* (NB) vs. normalized buffer size per port in the switch, for different T_{sw}, $n = 1$, $N = 8$; $\rho = 80\%$, $\rho_{uc} = 1\%$.

For $T_{sw} = B_{sw}$ (small thresholds), if B_{sw} is small, the cell loss rate for NSB as a function of B_{mx} saturates at rather high values, due to the fact that the throughput of the system is severely limited (*e.g.*, lower than 75% for $T_{sw} = B_{sw} = 4$, as shown in Section 3.1 above); for example, as shown in Fig. 5(a), the cell loss rate is always above 10^{-2} for $B_{sw} \leq 7$, and a B_{sw} as large as 10 is necessary to achieve low loss rates with reasonable values of B_{mx}. For $T_{sw} = NB_{sw}$ (large thresholds), once B_{sw} is sufficiently large, the slope of the curves of the cell loss rate for different B_{sw} is rather constant (see Fig. 5(b)). Due to the limited throughput in case of small T_{sw}, the total buffer requirements in the switch and multiplexers to achieve 10^{-6} cell loss rate with NSB are reasonable only for large T_{sw}, as shown in Fig. 6(a). For $T_{sw} < 4B_{sw}$, the total buffer requirements per port are even larger than those in a system with partitioned buffers in the switch with no backpressure (which, as shown in Fig. 7, would require 55 buffers per port to achieve the same loss rate). The buffer requirements wit NSB become comparable with those of partitioned buffers with no backpressure for $T_{sw} = 4B_{sw}$; in this case, for example, we could use 6 buffers per port in the switch and 52 buffers in the multiplexers to achieve the desired cell loss rate. The total buffer size is much larger than what needed in a fully-shared center-stage switch with no backpressure, where 11 buffers per port would suffice to achieve the same loss rate (see Fig. 7). In NSB, even with fully-shared center-stage buffers (*i.e.*, $T_{sw} = 8B_{sw}$), the total buffer requirements are smaller than those of partitioned buffers in the switch with no backpressure only if B_{sw} is large, in which case backpressure is only rarely applied (and the system approaches the case of fully-shared buffers and no backpressure); for example, to achieve 10^{-6} cell loss rate with NSB and $T_{sw} = 8B_{sw}$, we could use $B_{sw} = 5$, $B_{mx} = 45$, or $B_{sw} = 7$, $B_{mx} = 20$, or $B_{sw} = 9$, $B_{mx} = 10$.

The results for PPSB are dramatically different. In general, for the same B_{sw} and T_{sw}, PPSB requires significantly smaller buffers in the multiplexers than NSB to achieve a certain cell loss rate. With PPSB, very low cell loss rates can be obtained with small B_{sw}, for reasonable values of B_{mx}. For small T_{sw}, as shown in Fig. 5(c); the cell loss rate depends weakly on B_{sw}, since the center-stage buffer does not play much of a role if sharing is limited. For large T_{sw}, the curves of the cell loss rate become steeper and steeper as B_{sw} increases (see Fig. 5(d)). As far as the buffer requirements are concerned (see Fig. 6(b)), it is surprising to see that for small buffer sizes in the switch, partitioned buffers with PPSB require a significantly smaller number of buffers in the multiplexers than shared buffers to achieve a desired cell loss rate; for example, for $B_{sw} = 2$, $B_{mx} = 25$ is sufficient to achieve 10^{-6} loss rate with $T_{sw} = B_{sw}$, as opposed to a B_{mx} greater than 90 with $T_{sw} = 8B_{sw}$. For large buffers in the center stage, the impact of the center stage is more and more pronounced, and eventually shared buffers require less buffers in the multiplexer than partitioned buffers. The buffer requirements for $T_{sw} = B_{sw}$ and $T_{sw} = 2B_{sw}$ are roughly equivalent. The curves of the buffer requirements for small T_{sw} are rather flat, and become steeper for large T_{sw}, since the size of the center stage has more an impact on cell loss as sharing increases.

The buffer requirements of partitioned buffers in the center stage with PPSB are significantly lower than in the case of partitioned buffers with no backpressure. For example, to achieve 10^{-6} cell loss rate with PPSB we can use $B_{sw} = 2$ and $B_{mx} = 25$ (as opposed to $B_{sw} = 55$ without backpressure). The buffer requirements with PPSB are still larger than those of fully-shared buffers with no backpressure, but roughly equivalent to the case of shared buffers in the center stage with no backpressure and $T_{sw} = 2B_{sw}$.

These results indicate that the use of selective backpressure is actually a way to increase buffer utilization while using partitioned buffers. Using PPSB, the buffer capacity of the center stage can be expanded using the buffers in the multiplexers, paying only a relatively small penalty in buffer efficiency. In other words, although not quite as effective as "true" full-memory sharing, backpressure is a form of

"sharing". Furthermore, when fairness considerations dictate that a limitation on the sharing of the buffers in the center stage is used, the buffer requirements with backpressure become comparable with those of a shared center stage with no backpressure. The advantage of PPSB is that the high-speed buffers in the center-stage switch are relatively small, while most of the buffer capacity is concentrated in the relatively low-speed buffers in the multiplexers. This can translate into considerable cost advantages on the whole system with respect to a system with no backpressure. In the system without backpressure, where all the buffering is in the center-stage, it may be problematic to implement very large buffer capacities, due to the high speed of the buffers (and the problem is exacerbated as N increases); on the contrary, with selective backpressure, increasing the size of the low-speed buffers in the multiplexers is rather easily achieved.

4. Backpressure in Three-Stage Switch

In this section, we study the whole three-stage switch architecture described in Section 2.1 (see Fig. 1), with backpressure from the demultiplexers to the center-stage switch, and from the switch to the multiplexers. We consider NSB, PPSB, and PSSB. In the system under consideration, for each backpressure scheme, both throughput and loss rate depend on the complex interaction of a large number of switch parameters, such as B_{dx}, B_{sw}, B_{mx}, T_{sw}, N, and n. Clearly, an exhaustive study of all the combinations of these parameters would be prohibitive, and tedious to present. For this reason, we have considered a manageable number of specific cases, and studied them in depth. In particular, the results presented below are for $n = 4$. The buffers in each demultiplexer are fully shared among the queues corresponding to the subports; the buffers in each multiplexer are also fully shared. Finally, we assume $N = 8$ and $\rho_{wc} = 1\%$.

4.1. Throughput

We study the throughput of the switch model for normalized buffer sizes in each demultiplexer $B_{dx} = 1$, 20, and 100, and normalized buffer size per port in the switch $B_{sw} = 2$, 4, and 8. The traffic load on the center-stage switch is $\rho = 99\%$. To serve as a reference, in Fig. 8, we plot the throughput of a system without backpressure for the same sizes of B_{sw} and B_{dx}.

The throughput using NSB as a function of the normalized queue threshold in the switch T_{sw} is shown in Fig. 9 for $B_{mx} = 1000$ (*i.e.*, large buffers in the multiplexers), for various B_{sw} and B_{dx}. In all these plots, we recognize the familiar monotonic increase of the throughput with the switch threshold, since the less often backpressure is applied from the switch to the multiplexers, the less HOL blocking is introduced; for the same reason, the throughput also depends on the buffer size in the center-stage switch, although only for large thresholds. With small buffers in the demultiplexers (see Fig. 9(a), $B_{dx} = 1$), the throughput is severely limited (below 60%), even for relatively large buffer sizes in the switch. This heavy throughput degradation is therefore mainly due to starvation of the output subports, due to HOL blocking induced on cells stored in the center-stage switch by lack of buffer space in the demultiplexers. The throughput indeed improves significantly for the same switch size (see Fig. 9(d)) as the buffer size in the demultiplexers becomes large and HOL blocking is relieved. In any case, even for large buffers in the demultiplexers and in the center-stage switch, the maximum throughput remains well below 1 (as it

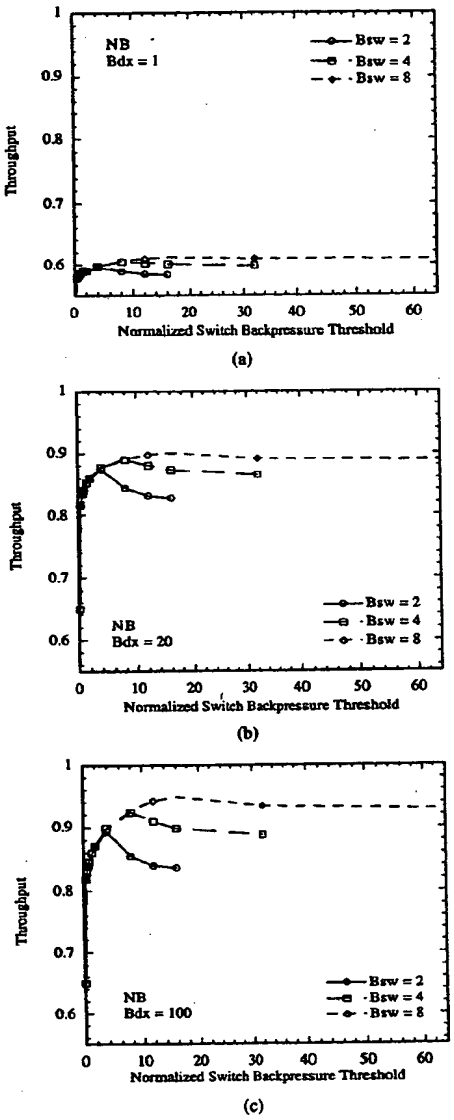

Fig. 8. Throughput of the three-stage switch architecture using *No Backpressure* (NB), as a function of the normalized queue threshold in the switch T_{sw}; $N = 8$, $n = 4$; $\rho = 99\%$, $\rho_{wc} = 1\%$; a) normalized buffer size per port in each demultiplexer $B_{dx} = 1$, for various normalized buffer sizes per port in the switch B_{sw}; b) $B_{dx} = 20$; c) $B_{dx} = 100$.

was the case in the two-stage switch with NSB). For small values of the switch threshold, the throughput is always severely limited, even for large sizes of the buffers in the demultiplexers (see Fig. 9(c), $B_{dx} = 100$, where the throughput is below 75% in the case of $T_{sw} = B_{sw} = 4$). Comparing Fig. 8 and Fig. 9, we observe that the throughput with NSB is in general lower than the maximum throughput that can be obtained in the system with no backpressure, for the same B_{sw} and B_{dx}, again confirming the fact that discarding cells is preferable, in terms of throughput, than storing them in the input multiplexers.

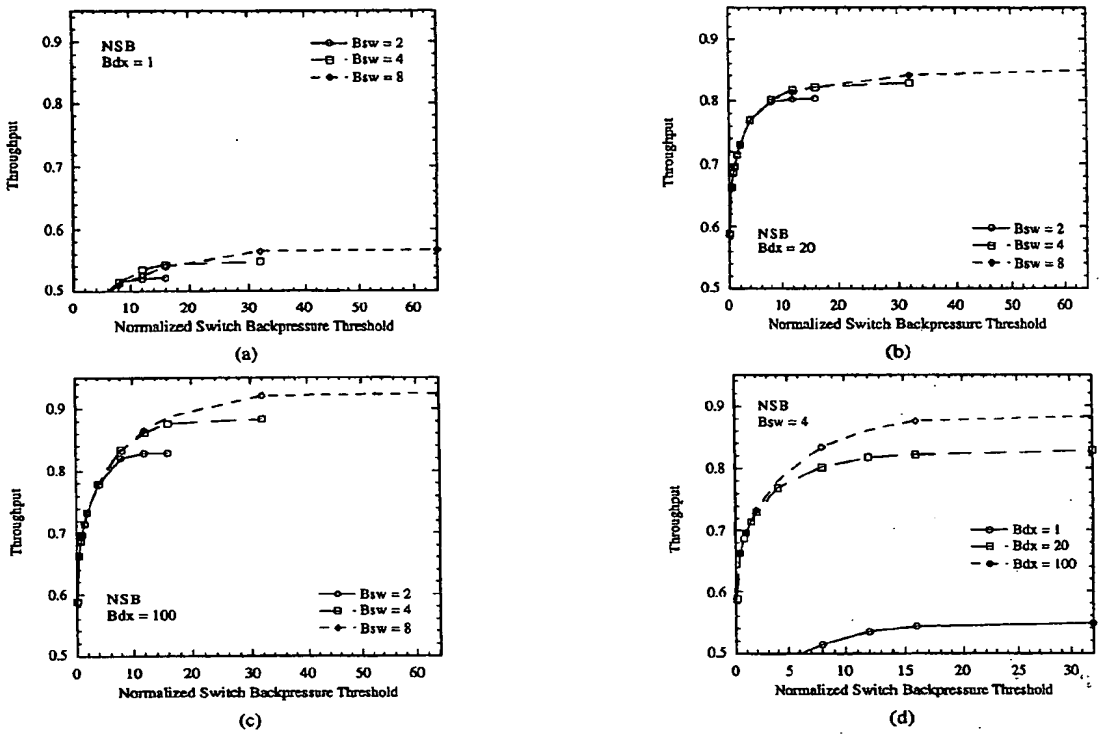

Fig. 9. Throughput of the three-stage switch architecture using NSB, as a function of the normalized queue threshold in the switch T_{sw}; normalized buffer size per port in each multiplexer $B_{mx} = 1000$; $N = 8$, $n = 4$; $\rho = 99\%$, $\rho_{sc} = 1\%$. a) Normalized buffer size per port in each demultiplexer $B_{dx} = 1$, for various normalized buffer sizes per port in the switch B_{sw}; b) $B_{dx} = 20$, for various B_{sw}; c) $B_{dx} = 100$, for various B_{sw}; d) $B_{sw} = 4$, for various B_{dx}.

In Fig. 10, we show the throughput using PPSB as a function of the normalized queue threshold in the switch T_{sw}, for a normalized buffer size in each multiplexer $B_{mx} = 1000$, for various B_{sw} and B_{dx}. For small sizes of the buffers in the demultiplexers (see Fig. 10(a), $B_{dx} = 1$), the throughput is still heavily limited, and only marginally better than with NSB. In fact, starvation of the output subports is still the dominant effect limiting the throughput, since the interaction between demultiplexers and center-stage switch is the same in PPSB and NSB. The throughput improves dramatically as the buffer size in the demultiplexer increases (*e.g.*, the maximum throughput is above 90% for $B_{dx} = 20$, and above 97% for $B_{dx} = 100$). As in the two-stage switch, for given buffer sizes in the demultiplexers and in the switch, the throughput is maximized for partitioned buffers in the center stage[6] (*i.e.*, $T_{sw} = B_{sw}$), and then degrades for large thresholds, as memory hogging becomes more and more likely (the degradation due to memory hogging is also noticeable in Fig. 8 in the system without backpressure). The degradation in throughput due to memory hogging, however, is rather marginal (equal to a few percentage points), especially for large switch sizes. As expected from the results of Section 3, the value of the throughput at the peak does not depend on the size of the buffer in the center-stage switch. This is because with PPSB, the buffers in the multiplexers effectively become an extension of the buffers in the switch (in particular in the case of partitioned buffers in the switch, corresponding to the peak in throughput), and the only degradation in throughput is due to HOL blocking induced by backpressure from the demultiplexers to the switch.

The throughput using PSSB as a function of the normalized queue threshold in the switch T_{sw}, for $B_{mx} = 1000$, for various B_{sw} and B_{dx}, is shown in Fig. 11. In this case, *no* HOL blocking is introduced in the system. As a consequence, the maximum throughput does not depend on the size of the buffer in the demultiplexer and in the switch, and is always equal to the offered load. The curves for PSSB have a similar trend as those for PPSB. The throughput peaks for the case of partitioned buffers in the center-stage switch, and then degrades due to the memory-hogging effect; it is important to remember that in PSSB there are Nn queues in the center stage switch, so the case of partitioned buffers occurs for $T_{sw} = B_{sw}/n = B_{sw}/4$. Similarly to PPSB, the degradation due to memory hogging is marginal, and decreases as the buffer size in the switch increases. Note that, although the maximum throughput is independent from B_{dx}, the worst-case throughput for a given B_{sw} (corresponding to fully shared-buffers, $T_{sw} = NB_{sw} = 8B_{sw}$) decreases with the buffer size in the demultiplexer (see Fig. 11(d)); this is because the smaller B_{dx} is, the more the traffic

[6] More precisely, in case of small sizes of the demultiplexer's buffer, the maximum occurs for threshold values slightly larger than the value for partitioned buffers (see the curve for $B_{sw} = 8$ in Fig. 10(a), for which the maximum throughput occurs for $T_{sw} = 12$, rather than $T_{sw} = 8$). This is due to the effect of increased sharing which, in case of a large buffer size in the switch, visibly relieves the heavy HOL blocking induced by the small buffer size in the demultiplexer; thus, in this particular situation, it dominates the degradation due to memory hogging, for a small range of threshold values.

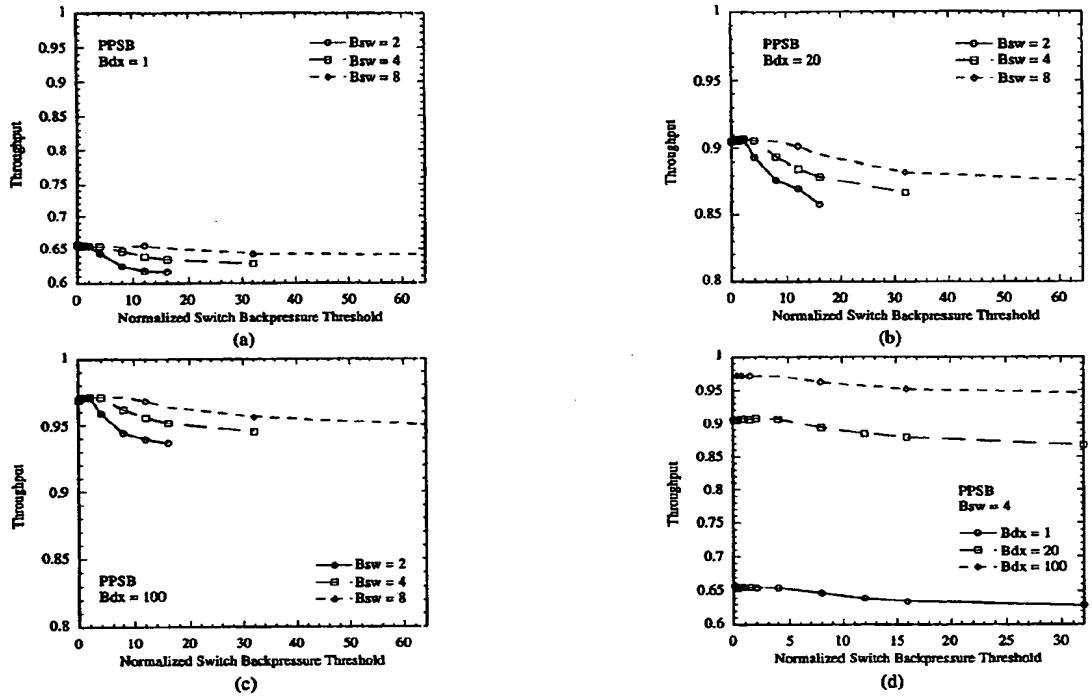

Fig. 10. Throughput of the three-stage switch architecture using PPSB, as a function of T_{sw}; $B_{mx} = 1000$; $N = 8$, n = 4; $\rho = 99\%$, $\rho_{uc} = 1\%$. a) $B_{dx} = 1$, for various B_{sw}; b) $B_{dx} = 20$, for various B_{sw}; c) $B_{dx} = 100$, for various B_{sw}; d) $B_{sw} = 4$, for various B_{dx}.

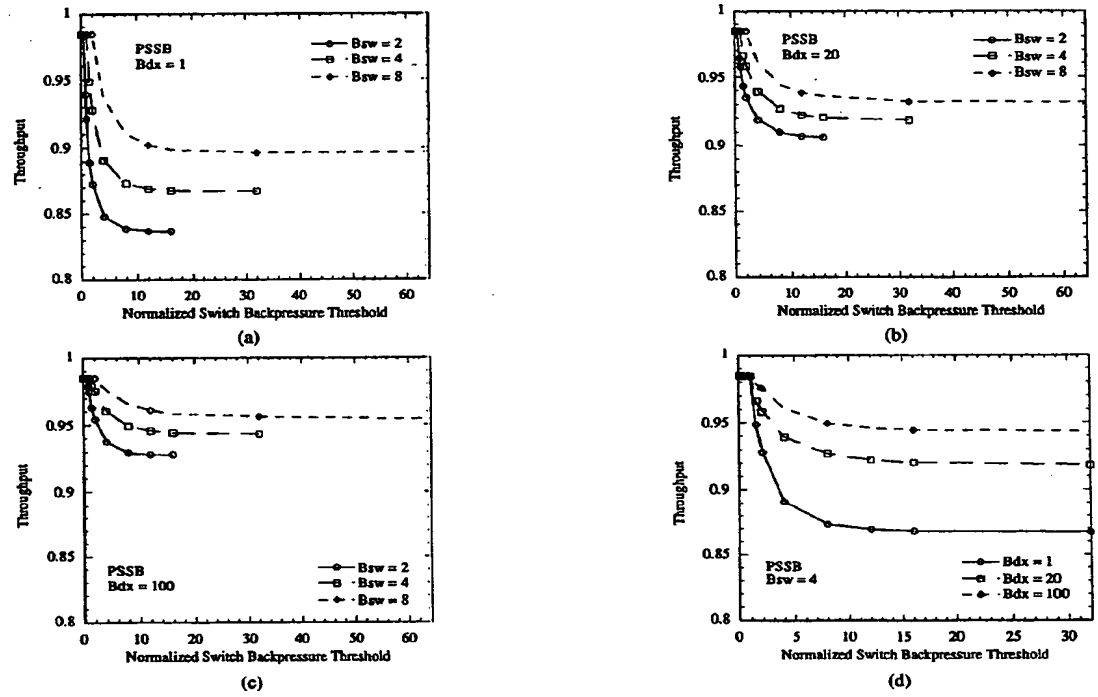

Fig. 11. Throughput of the three-stage switch architecture using PSSB, as a function of T_{sw}; $B_{mx} = 1000$; $N = 8$, n = 4; $\rho = 99\%$, $\rho_{uc} = 1\%$. a) $B_{dx} = 1$, for various B_{sw}; b) $B_{dx} = 20$, for various B_{sw}; c) $B_{dx} = 100$, for various B_{sw}; d) $B_{sw} = 4$, for various B_{dx}.

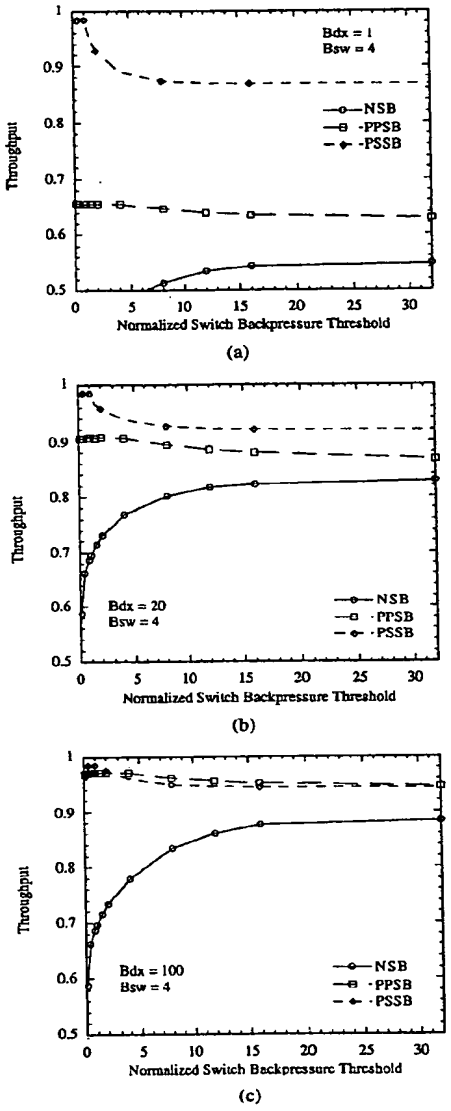

Fig. 12. Throughput of the three-stage switch architecture using NSB, PPSB, and PSSB as a function of the normalized queue threshold in the switch T_{sw}; $B_{mx} = 1000$; $B_{sw} = 4$; $N = 8$, n = 4; $\rho = 99\%$, $\rho_{vc} = 1\%$. a) $B_{dx} = 1$; b) $B_{dx} = 20$; c) $B_{dx} = 100$.

is pushed back in the first two stages in the system, the more important is that the buffer in the switch is not momentarily hogged.

The throughput for the three backpressure schemes as a function of T_{sw} is compared in Fig. 12 for various B_{dx}, $B_{sw} = 4$, $B_{mx} = 1000$. For small B_{dx} (see Fig. 12(a), $B_{dx} = 1$), PSSB achieves 99% throughput with partitioned buffers ($T_{sw} = B_{sw}/n$) in the center stage, but the throughput decreases rapidly as the threshold increases and saturates just below 90% for $T_{sw} \geq B_{sw}$. The other two schemes are severely limited by HOL blocking. PPSB performs visibly better than NSB; it reaches a maximum throughput of about 65% for partitioned buffers and slightly decreases for larger thresholds. The throughput for NSB is very low (50%) for small threshold values, and increases monotonically up to 54%. For larger buffers in the demultiplexers, as expected, the differences in the maximum throughput for the three scheme are less dramatic, since backpressure is applied less and less often. For $B_{dx} = 20$ (see Fig. 12(b)), NSB still shows heavy throughput degradation for small thresholds, and a maximum throughput of about 82%. PPSB has a maximum throughput of about 91%, and its throughput is always above 87%. For PSSB, the throughput peaks at 99% and stays above 92%.

It is interesting to note that for $B_{dx} = 100$ (see Fig. 12(c)), PPSB and PSSB have roughly comparable throughputs (although PPSB never reaches 99% throughput). For $T_{sw} > B_{sw}$, PPSB actually performs *better* than PSSB; this is due to the fact that throughput degradation due to memory hogging is more significant as the number of queues in the center stage increases. NSB still shows heavy degradation for small thresholds and saturates below 90%.

Finally, PPSB and PSSB retain their advantage in throughput over NSB even in the case of fully-shared buffers, when we would expect all schemes to look very similar. As noted in Section 3, this is due to the smoothing effect in the traffic offered to the center stage introduced by the round-robin service of the queues in the multiplexers. Similarly, PSSB has an advantage over PPSB also for fully shared buffers (except in the case of large buffers in the demultiplexers), due to the smoothing effect in the traffic offered to the demultiplexers introduced by the finer round-robin service in the multiplexers, and by the round-robin service of the different queues corresponding to each output port in the center-stage switch.

To summarize this discussion, we have shown that, in case of PSSB, the buffers in the input multiplexers are used effectively as an extension of the buffers in the demultiplexers without introducing blocking, so that a small buffer size in the demultiplexers is sufficient to achieve 99% throughput (in other words, the buffers in the demultiplexers have been "relocated" into the center-stage switch and into the multiplexers). With PPSB, the buffers in the multiplexers are used effectively as an extension of the buffers in the center-stage switch, but can only be partially considered as an extension of the buffers in the demultiplexers, due to HOL blocking. Thus, in order to achieve high throughput, large buffers in the demultiplexers must be used; once the buffers in the demultiplexers are large, PPSB is essentially equivalent to PSSB. In NSB, because of HOL blocking, the buffers in the multiplexers are not an extension of the buffers in the switch and demultiplexers, and throughput degradation results.

4.2. Cell Loss Rate

In this section, we study the cell loss rate in the three-stage switch. As we said above, we have selected $n = 4$, and assumed that the buffers in the multiplexers and demultiplexers are fully shared. The results presented here are for a load $\rho = 80\%$, $\rho_{vc} = 1\%$.

The cell loss rate vs. the normalized queue threshold in the switch T_{sw}, for different normalized buffer sizes B_{dx}, B_{sw}, and B_{mx}, using NSB, is presented in Fig. 13. As it was the case for the throughput, for given buffer sizes, the lowest loss rate is obtained for fully-shared buffers in the center-stage switch; in general, the loss rate is very sensitive to the value of the threshold. The cell loss rate depends on the value of B_{mx}, especially in case of fully-shared buffers. For $B_{sw} = 4$, $B_{dx} = 20$,

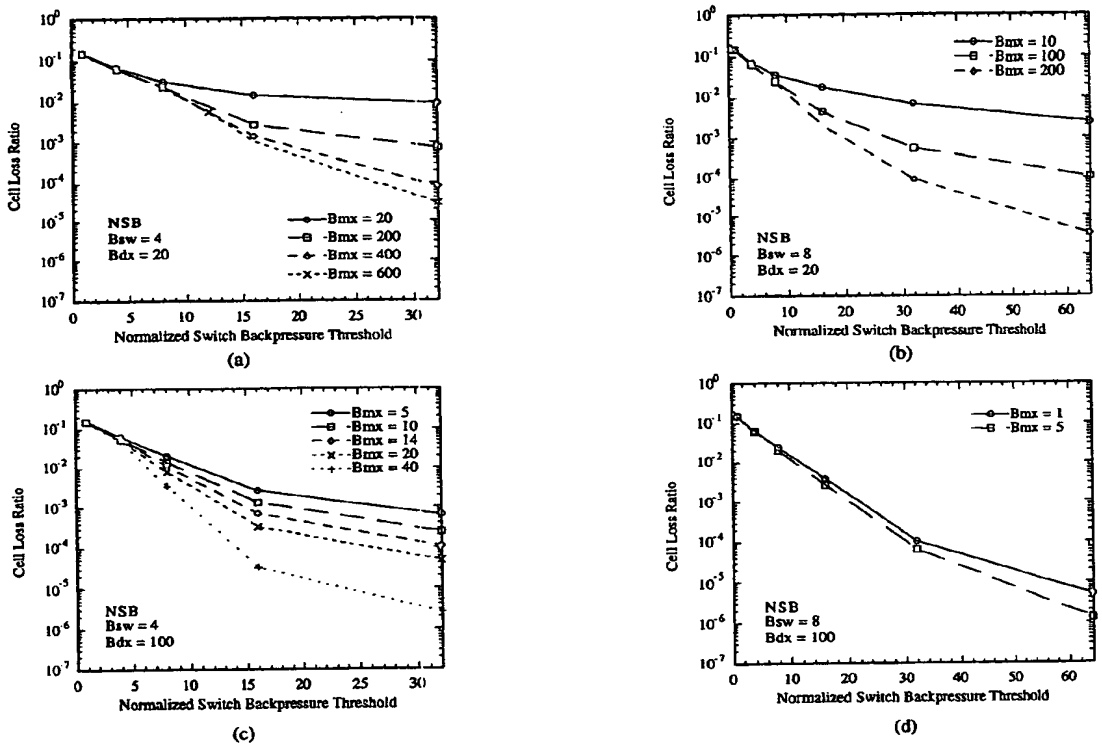

Fig. 13. Cell loss rate using NSB, as a function of the normalized queue threshold in the switch T_{sw}, for different normalized buffer sizes per port in each multiplexer B_{mx}; $N = 8$, $n = 4$; $\rho = 80\%$;$\rho_{sw} = 1\%$. a) Normalized buffer sizes per port in the switch $B_{sw} = 4$, normalized buffer size per port in each demultiplexer $B_{dx} = 20$; b) $B_{sw} = 8$, $B_{dx} = 20$; c) $B_{sw} = 4$, $B_{dx} = 100$; d) $B_{sw} = 8$, $B_{dx} = 100$.

very large buffers are required in the multiplexers to achieve low loss rates, as shown in Fig. 13(a); for example, $B_{mx} = 600$ yield a minimum loss of 10^{-4}; the loss improves very slowly as the buffer size in the multiplexers increases, and eventually saturates at relatively large values. A visible improvement is obtained for $B_{sw} = 8$, as shown in Fig. 13(b) (*e.g.*, $B_{mx} = 200$ yields a cell loss rate below 10^{-5}). For $B_{dx} = 100$, much smaller buffers are needed in the multiplexers. For example, with $B_{sw} = 4$, $B_{mx} = 20$ (as opposed to 600 in the case of $B_{dx} = 20$) achieves 10^{-4} loss rate, and the loss rate keeps improving with B_{mx}; for $B_{mx} = 8$, $B_{mx} = 1$ yields a cell loss rate lower than 10^{-5}, and comparable to the one obtained with $B_{mx} = 200$ in the case of $B_{dx} = 20$. For partitioned buffers in the switch, the system suffers high losses even for large buffers, and the loss does not depend on B_{mx} and B_{dx}, since it is dominated by insufficient buffer in the center stage. In general, reasonable buffer requirements are obtained only by using relatively large buffers in the demultiplexers and in the center-stage switch, so that backpressure is applied rather rarely, and the system approaches the case without backpressure; furthermore, the buffers in the switch have to be fully shared.

In Fig. 14, we show the cell loss rate for PPSB as a function of the normalized queue threshold in the switch T_{sw}, for different B_{dx}, B_{sw}, and B_{mx}. For $B_{mx} = 4$, the curves of the cell loss rate show a minimum for small T_{sw}, but for a value of the threshold larger than $T_{sw} = B_{mx}$ (*i.e.*, partitioned buffers). This is due to the two contrasting effects of increased sharing and increased chance of memory hogging as the threshold increases. As B_{mx} increases, the minimum for the cell loss rate occurs for larger values of T_{sw} (compare Fig. 14(a) and 14(d)); in fact, for larger buffer sizes, the beneficial effect of sharing dominates more and more over the chance of memory hogging. For this reason, for both large B_{sw} and B_{dx}, the minimum of the cell loss rate eventually occurs for fully-shared buffers (see Fig. 14(d)). This behavior is quite different from the one of the throughput curves. For $B_{sw} = 4$ and $B_{dx} = 20$ (see Fig. 14(a)), $B_{mx} = 30$ gives a minimum cell loss rate of 10^{-4} (compare this with B_{mx} larger than 600 in case of NSB to achieve the same cell loss rate). With $B_{sw} = 8$, $B_{mx} = 14$, as shown in Fig. 14(b), achieves a cell loss rate below 10^{-3}. For $B_{sw} = 4$ and $B_{dx} = 100$, the same $B_{mx} = 14$ yields a cell loss rate below 10^{-5} (see Fig. 14(c)). For $B_{sw} = 8$ and $B_{dx} = 100$, $B_{mx} = 5$ gives a loss rate below 10^{-6}, as depicted in Fig. 14(d).

As expected, the difference between NSB and PPSB in buffer size required in the multiplexers to achieve a certain loss rate decreases as B_{dx} and B_{sw} increase, since backpressure is less and less of a factor as the switch size increases. For given buffer sizes, however, the difference in cell loss rate between the two schemes is always significant; the curves of PPSB are always below the corresponding curves of NSB. This holds even for fully-shared buffers, and even for large buffer sizes, due to the effect, mentioned above, of smoothing in the traffic pattern offered to the switch introduced by the round-robin service of the multiplexers' queues used in PPSB.

The cell loss rate using PSSB vs. the normalized queue threshold in the switch T_{sw}, for different B_{dx}, B_{sw}, and B_{mx}, is depicted in Fig. 15. The curves of the cell loss rate for PSSB have a trend similar to those

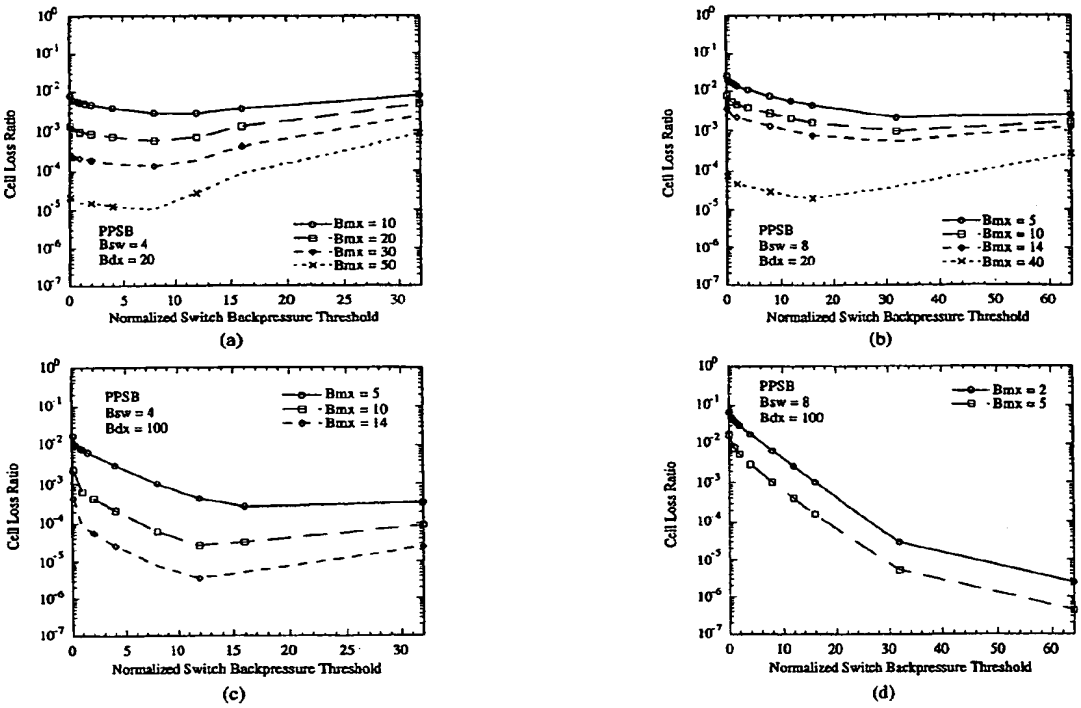

Fig. 14. Cell loss rate using PPSB, as a function of T_{sw}, for different B_{mx}; $N = 8$, n = 4; $\rho = 80\%$;$\rho_{vc} = 1\%$. a) $B_{sw} = 4$, $B_{dx} = 20$; b) $B_{sw} = 8$, $B_{dx} = 20$; c) $B_{sw} = 4$, $B_{dx} = 100$; d) $B_{sw} = 8$, $B_{dx} = 100$.

for PPSB; in case of small B_{sw}, they show a minimum for small thresholds (for a value of T_{sw} generally larger than $T_{sw} = B_{mx}/n$, *i.e.*, the case of partitioned buffers), and the minimum occurs for higher values of T_{sw} as B_{sw} increases, until occurs for fully-shared buffers in case of large B_{sw} and B_{dx}.

In case of small buffers in the demultiplexers, PSSB performs significantly better than the other two schemes; this is intuitively expected from the results discussed in the previous section, since PSSB is the only scheme which uses the buffer space in the first two stages effectively as an extension of the buffers in the demultiplexers. For example, for $B_{sw} = 4$ and $B_{dx} = 20$, $B_{mx} = 20$ gives a minimum cell loss rate below 10^{-5} (see Fig. 15(a)), as opposed to a cell loss rate just below 10^{-3} in case of PPSB. From Fig. 15(b), we see that for $B_{sw} = 8$, $B_{mx} = 14$ is sufficient to achieve 10^{-6} minimum loss rate (as opposed to a loss rate of about 10^{-3} in case of PPSB).

For larger sizes of the buffers in the demultiplexers, the advantage of PSSB over PPSB almost disappears. In case of $B_{sw} = 4$, $B_{dx} = 100$, $B_{mx} = 14$ gives a minimum loss rate of 10^{-6} (see Fig. 15(c)). The cell loss rate at the minimum is only marginally smaller than the corresponding minimum cell loss rate with PPSB; for $T_{sw} \geq 10$, the corresponding curve for PPSB is actually *below* the curve for PSSB; this is due to the smaller number of queues in the center stage in case of PPSB. For the same reason, in case of $B_{sw} = 8$, $B_{dx} = 100$, the minimum loss for PPSB (which occurs for fully-shared buffers) is actually *lower* than the minimum loss for PSSB (see Fig. 15(d)); for smaller T_{sw}, however, the loss is smaller in the case of PSSB (the minimum for PSSB occurs for $T_{sw} \geq 32$).

5. Conclusions

In this paper, we have studied different schemes to implement backpressure in a system consisting of a center-stage shared-memory switch with input multiplexers and output demultiplexers. We have shown that *Non-Selective Backpressure* (NSB) introduces heavy HOL blocking, causing throughput degradation; in terms of total buffer requirements, NSB does not offer any advantages even respect to a system with completely-partitioned buffers and no backpressure, and still requires large buffers in the center stage and in the demultiplexers. On the contrary, both *Per-Port Selective Backpressure* (PPSB) and *Per-Subport Selective Backpressure* (PSSB) can achieve high throughput; they have buffer requirements that are much smaller than those of partitioned buffers without backpressure, and not very far from those of a fully-shared center stage without backpressure. In other words, PPSB and PSSB are two ways to achieve higher buffer utilization than partitioned buffers, while keeping the majority of the buffer capacity physically separate in the input multiplexers. In terms of throughput, the two schemes perform best with partitioned buffer; in terms of loss rate, some sharing of the center-stage buffers is beneficial; however, both schemes behave well with buffers that are not fully shared, as it may be required to guarantee fairness in the switch.

If the buffers in the demultiplexers need to be small, then PSSB is the only choice to achieve high throughput and low cell loss rates with reasonable buffer sizes. However, if the buffers in the demultiplexers are sufficiently large, then the performance of PPSB

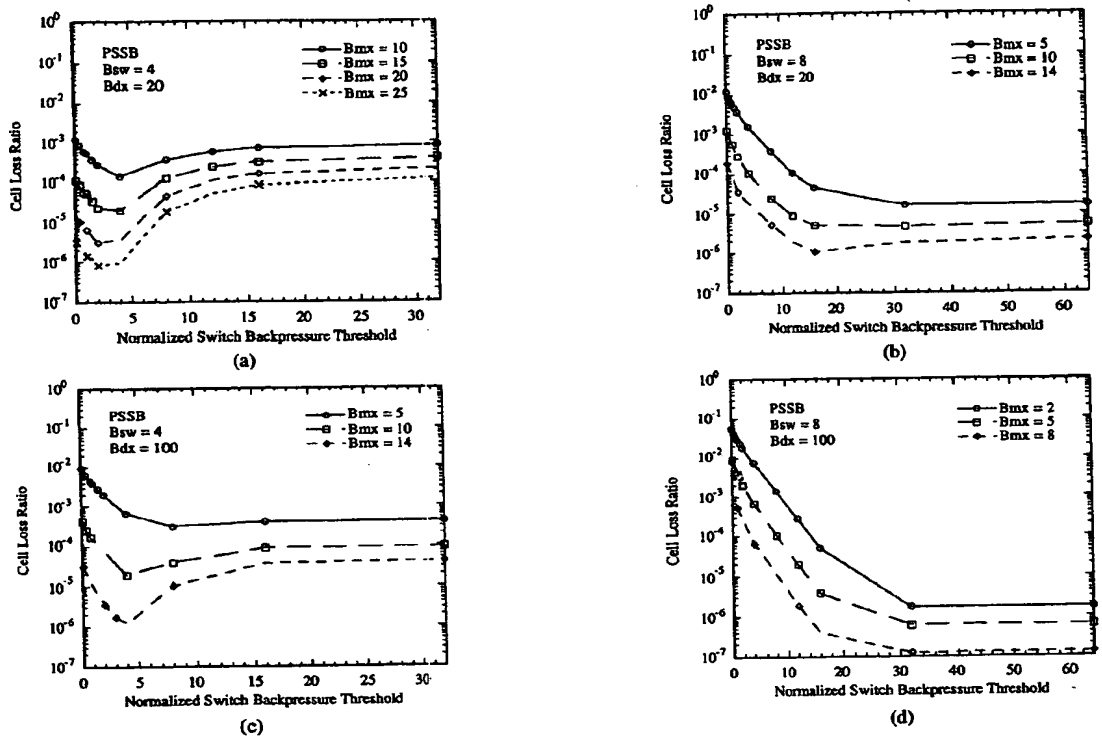

Fig. 15. Cell loss rate using PSSB, as a function of T_{sw}, for different B_{sx}; $N = 8$, n = 4; $\rho = 80\%$;$\rho_{vc} = 1\%$. a) $B_{sw} = 4$, $B_{dx} = 20$; b) $B_{sw} = 8$, $B_{dx} = 20$; c) $B_{sw} = 4$, $B_{dx} = 100$; d) $B_{sw} = 8$, $B_{dx} = 100$.

and PSSB are comparable. From this perspective, PPSB is therefore a simple way to expand the buffer capacity of the center stage without introducing HOL blocking; the buffers in the demultiplexers can then be sized sufficiently large to perform the demultiplexing function, as it is necessary in a system without backpressure.

6. References

[1] T. Kozaki *et al.*, "32x32 Shared Buffer Type ATM Switch VLSI's for B-ISDN's," *IEEE Jour. Select. Areas Comm.*, pp. 1239-1247, Oct. 1991.

[2] Y. Shobatake *et al.*, "A One-Chip Scalable 8x8 ATM Switch LSI Employing Shared Buffer Architecture," *IEEE Jour. Select. Areas Comm.*, pp. 1248-1254, Oct. 1991.

[3] H. Kuwahara *et al.*, "A Shared Buffer Memory Switch for an ATM Exchange," *Proc. ICC '89*, paper 4.4, Boston, MA, June 1989.

[4] A. E. Eckberg and T.-C. Hou, "Effects of Output Buffer Sharing on Buffer Requirements in an ATDM Packet Switch," *Proc. INFOCOM'88*, paper 5A.4, New Orleans, LA, March 1988.

[5] N. Endo *et al.*, "Traffic Characteristics Evaluation of a Shared Buffer ATM Switch," *Proc. GLOBECOM '90*, paper 905.1, San Diego, CA, Dec. 1990.

[6] M. J. Karol and K. Y. Eng, "Performance of Hierarchical Multiplexing in ATM Switch Design," *Proc. ICC'92*, pp. 269-275, Chicago, IL, June 1992.

[7] I. Iliadis and W. E. Denzel, "Performance of Packet Switches with Input and Output Queueing," *Proc. ICC'90*, paper 316.3, Atlanta, GA, April 1990.

[8] I. Iliadis and W. E. Denzel, "Analysis of Packet Switches with Input and Output Queueing," *IEEE Trans. Comm.*, 41, pp. 731-740, May 1993.

[9] A. K. Gupta and N. D. Georganas, "Analysis of a Packet Switch with Input and Output Buffers and Speed Constraints," *Proc. INFOCOM'91*, paper 7A.2, Bal Harbour, FL, April 1991.

[10] G. Bruzzi and A. Pattavina, "Performance Evaluation of an Input-Queued ATM Switch with Internal Speed-up and Finite Output Queues," *Proc. GLOBECOM'90*, paper 801.5, San Diego, CA, Dec. 1990.

[11] G. Bruzzi and A. Pattavina, "Analysis of Input and Output Queueing for Nonblocking ATM Switches," *IEEE/ACM Trans. Networking*, 1, pp. 314-328, June 1993.

[12] T. D. Morris ad H. G. Perros, "Performance Modeling of a Multi-Buffered Banyan Switch Under Bursty Traffic," *Proc. INFOCOM'92*, pp. 436-445, May 1992.

[13] A. I. Elwalid and I. Widjaja, "Efficient Analysis of Buffered Multistage Networks under Bursty Traffic," *Proc. GLOBECOM'93*, pp. 1067-1071, Houston, TX, Nov. 1993.

[14] D. Basak, A. K. Choudhury and E. L. Hahne, "Sharing Memory in Multistage ATM Switches", *Proc. Fourth Int. Conf. Computer Communications and Networks*, Las Vegas, Nevada, Sept. 1995.

[15] A. K. Choudhury and E. L. Hahne, "Buffer Management in A Hierarchical Shared Memory Switch", *Proc. INFOCOM'94*, pp. 1410-1419, Toronto, Canada, June 1994.

[16] F. M. Chiussi, "Design, Performance, and Implementation of a Three-Stage Banyan-Based Architecture with Input and Output Buffers for Large Fast Packet Switches," *Tech. Rep. No. CSL-TR-93-573*, Stanford University, Stanford, CA, June 1993.

[17] A. K. Choudhury and E. L. Hahne, "Dynamic Queue Length Thresholds in a Shared Memory ATM Switch," *Proc. INFOCOM'96*, San Francisco, CA, March 1996.

[18] A. R. Bonde and S. Ghosh, "A Comparative Study of Fuzzy Versus Fixed Thresholds for Robust Queue Management in Cell-Switching Networks," *IEEE/ACM Trans. Networking*, 2, pp. 337-344, Aug. 1994.

[19] R. O. LaMaire and D. N. Serpanos, "Two-Dimensional Round-Robin Schedules for Packet Switches with Multiple Input Queues," *IEEE/ACM Trans. Networking*, 2, pp. 471-482, Oct. 1994.